Islas Canarias
Del 15 al 19 de noviembre de 2021

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

XXI Jornadas de Estadística de las Comunidades Autónomas

VALIDACIÓN DE FICHEROS ADMINISTRATIVOS DE CARÁCTER ECONÓMICO

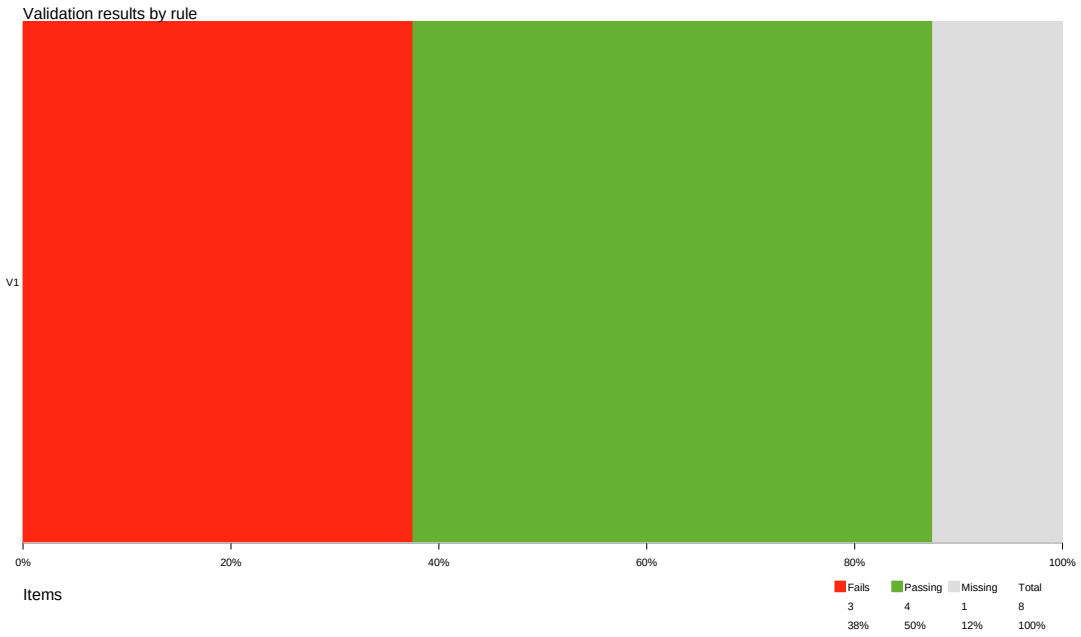
Matias Ledesma Sereno
Ander Iparraguirre García

Matias Ledesma Sereno: https://www.linkedin.com/in/matiasledesmasereno/
Ander Iparraguirre García: https://www.linkedin.com/in/ander-iparraguirre/

1. Validación

- "La validación consiste en verificar si un valor o una combinación de valores es aceptable en un conjunto de combinaciones" (Di Zio, M. et al, 2018)

```
library(validate)
a_df <- data.frame(a = c (3,5,1,-4,7,23,-18,NA))
rules <- validator(a<=5 & a>=-4)
plot(confront(a_df, rules))
```



2. Objetivo

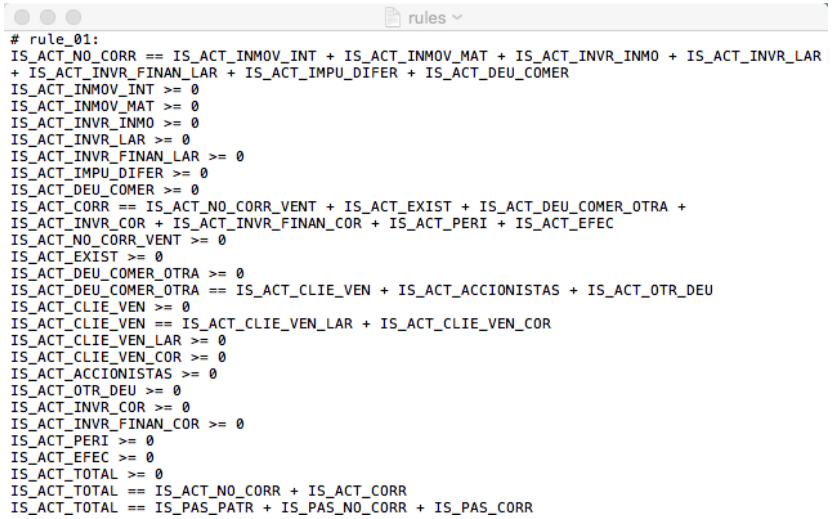
- El objetivo de este trabajo ha sido establecer un proceso estándar de validación para ficheros de carácter económico, bien provenientes de encuestas o bien de registros administrativos.
- El fichero que se usará para este ejemplo será el del "Impuesto de Sociedades 2018".

3. Metodología

- El sistema estadístico Europeo (ESS) creó un grupo para producir datos de calidad para la UE. ESSnet Validat es el proyecto que ha construido un entorno de validación con la ambición de sistematizar el proceso de validación.
- Se procederá a la validación de registros administrativos mediante reglas de carácter contable. Para ello se usarán librerías del repositorio "Data cleaning for statistical purpose" creado por Mark Van der Loo y Edwin de Jong.
- Este repositorio incluye gran variedad de librerías las cuales tienen el propósito de validar y limpiar ficheros.

3.1 Reglas

- Para proceder con la evaluación es necesario confrontar una serie de reglas con el fichero a validar. Estas reglas deben de almacenarse en un fichero .txt o pueden ser ejecutadas con la función "validator".
- En dicho caso se han aplicado reglas de carácter contable, pero existen gran variedad de reglas:
  - Formato
  - Rangos numéricos
  - Rangos temporales
  - Condiciones
  - ...



3.2 Validatetools

- Validatetools sirve para gestionar y evaluar los sets de reglas.

```
library(validatetools)
# Se construye un objeto S4 con las reglas pertinentes
rules <- validator(V1 = a > 0, V2 = b > 0, V3 = a + b <0)
# Se detecta si el set de reglas es no factible
is_infeasible(rules)
> [1] TRUE
# Detecta las reglas que no son factibles
detect_infeasible_rules(rules)
> [1] "V1"
```

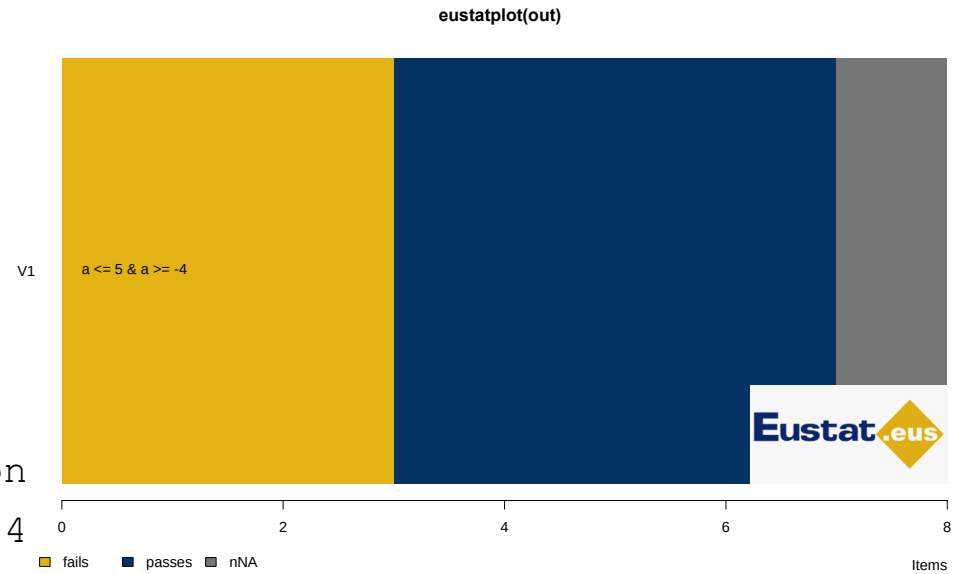
3.2 Validatetools

```
# Aconseja de qué manera sería factible el set de reglas
# creando un nuevo set de reglas new_rules <-
# make_feasible(rules)
make_feasible(rules)
> Dropping rule(s): "V1"
> Object of class 'validator' with 2 elements:
> V2: b > 0
> V3: a + b < 0
> Rules are evaluated using locally defined options
```

3.3 Validate

- Esta es la librería principal del entorno de "Data Cleaning for Statistical Purpose". Con ella se evalúa la calidad de los registros según reglas. También existe la versión que valida bases de datos (validatedb).

```
library(validate)
a_df <- data.frame(a = c (3,5,1,-4,7,23,-18,NA))
rules <- validator(a<=5 & a>=-4)
out <- confront(a_df, rules)
summary(out)
>  name items passes fails nNA error warning       expression
>    V1     8      4     3   1 FALSE   FALSE a <= 5 & a >= -4
```



3.4 Errorlocate

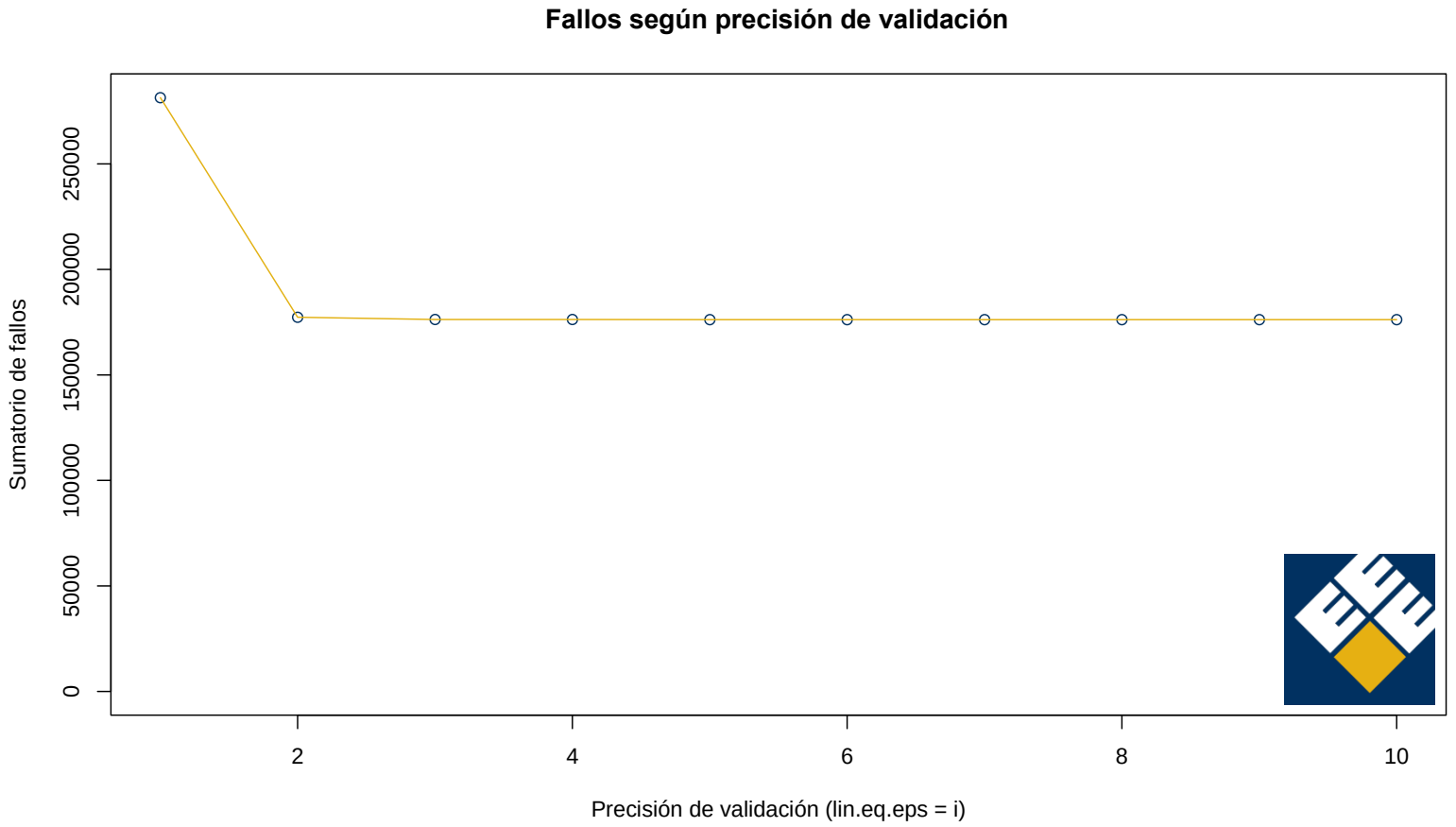
- Se dedica a encontrar el menor número de campos a alterar posibles para que se cumplan el mayor número de reglas usando el algoritmo de Felligi-Holt.
- Traduce los datos y las reglas para generar un problema de programación lineal de enteros mixtos.
- Consume muchos recursos, por lo que mandar un "subsets" es óptimo.

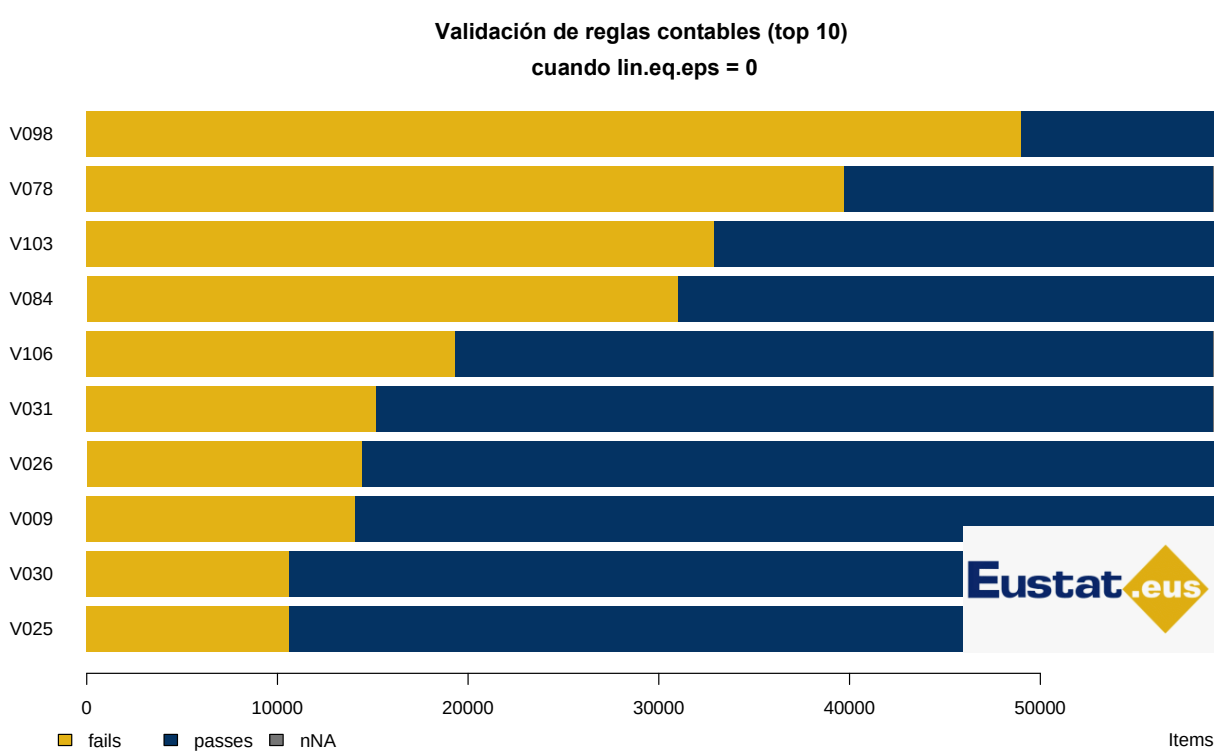
```
a_df <- data.frame(a = c (3,5,1),
                   b = c (4,9,-11),
                   c = c (7, 40, 8))
rules <- validator(b>0, a+b==c)
out <- locate-error(a_df, rules)
# param: "timeout" restringe el tiempo usado por registro: "Ncpus" sirve para
# seleccionar el número de núcleos usados. También se puede añadir pesos a las variables.
```

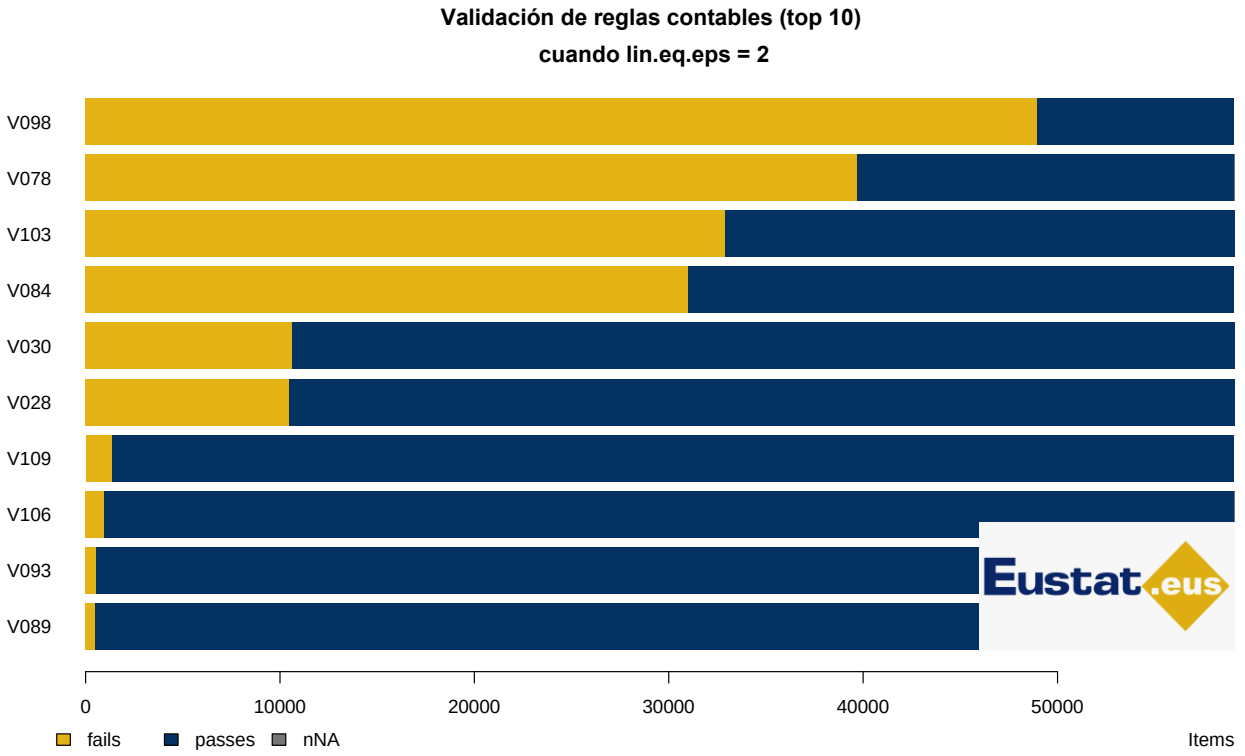
3.4 Errorlocate

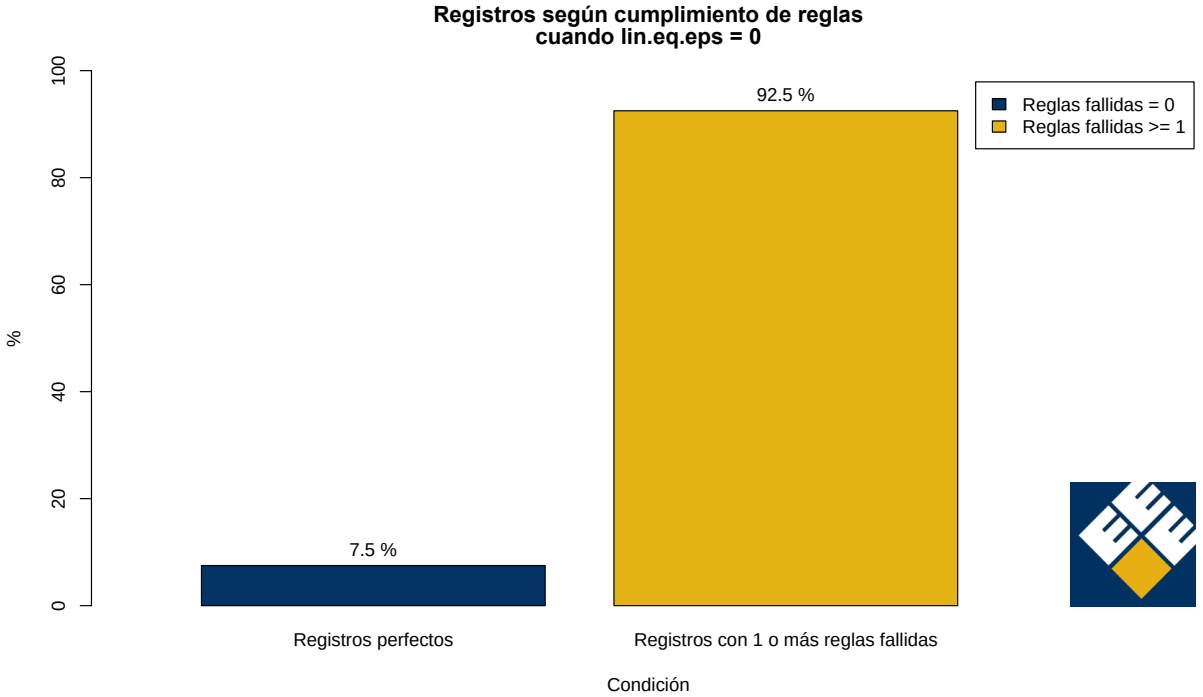
```
summary(out)
#> Variable:
#>  name errors missing
#>1    a      1       0
#>2    b      1       0
#>3    c      0       0
#>Errors per record:
#>  errors records
#>1      0       1
#>2      1       2
# La siguiente función pretende sustituir los  # campos erróneos con NA para        # proceder a una posterior imputación
b_df <- replace_errors(a_df, out, value = "NA")
#>   a  b  c
#>1  3  4  7
#>2 NA  9 40
#>3  1 NA  8
```

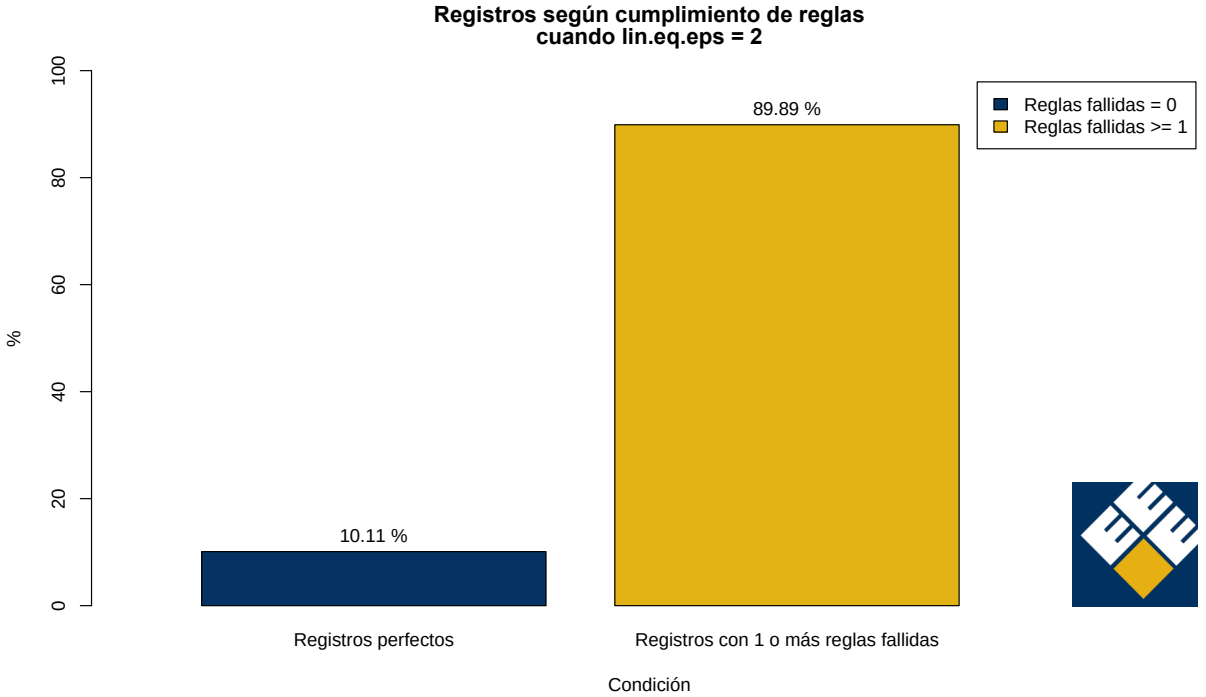
4. Resultados

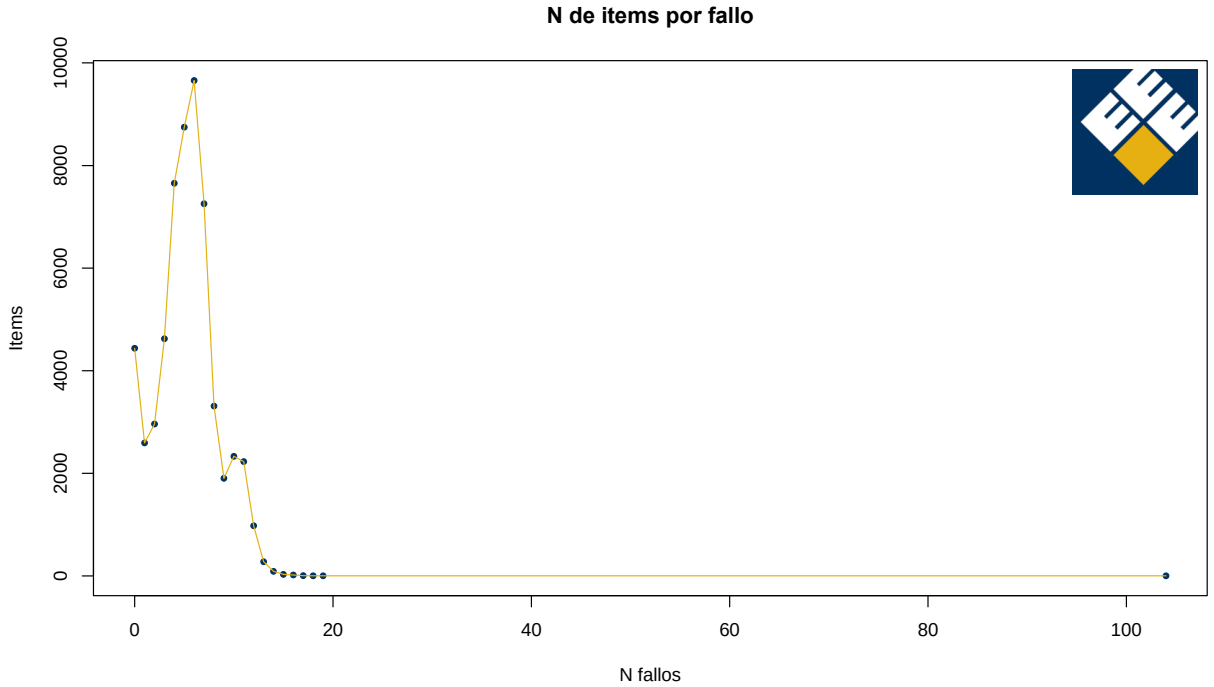


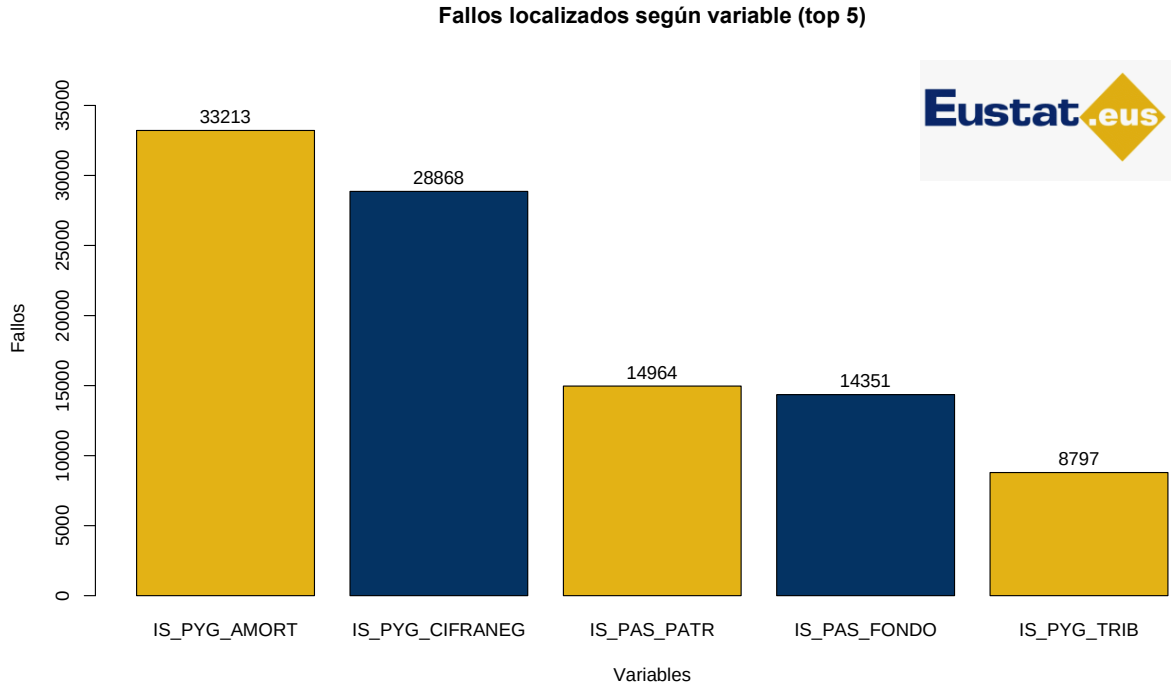

Validación de reglas contables (top 10)
cuando lin.eq.eps = 0
V098
V078
V103
V084
V106
V031
V026
V009
V030
V025
0
10000
20000
30000
40000
50000
fails
passes
nNA
Items
Eustat.eus


Validación de reglas contables (top 10)
cuando lin.eq.eps = 2
V098
V078
V103
V084
V030
V028
V109
V106
V093
V089
0
10000
20000
30000
40000
50000
fails
passes
nNA
Items
Eustat.eus


Registros según cumplimiento de reglas
cuando lin.eq.eps = 0
%
0
20
40
60
80
100
7.5 %
92.5 %
Reglas fallidas = 0
Reglas fallidas >= 1
Registros perfectos
Registros con 1 o más reglas fallidas
Condición


Registros según cumplimiento de reglas
cuando lin.eq.eps = 2
%
0
20
40
60
80
100
10.11 %
89.89 %
Reglas fallidas = 0
Reglas fallidas >= 1
Registros perfectos
Registros con 1 o más reglas fallidas
Condición


N de items por fallo
Items
0
2000
4000
6000
8000
10000
0
20
40
60
80
100
N fallos


Fallos localizados según variable (top 5)
Fallos
0
5000
10000
15000
20000
25000
30000
35000
33213
28868
14964
14351
8797
IS_PYG_AMORT
IS_PYG_CIFRANEG
IS_PAS_PATR
IS_PAS_FONDO
IS_PYG_TRIB
Variables

5. Conclusiones

- Software útil para automatizar el proceso de validación.
- Generar un informe de calidad del fichero en base a un set de reglas (mediante RMarkdown). Validatereport es una alternativa que exporta resultados de validate a un informe JSON estándar de ESSnet.
- ValidatReport es una herramienta que permite efectuar tareas de validación con interfaz gráfica mediante shiny.
- Problema: inestabilidad cuando los coeficientes de una matriz difieren de 10e10 al usar lpSolveAPI.
  - Reescalar el fichero para ejecutar funciones como locate_errors, replace_errors...

6. Bibliografía

- Bantilan, N. (2020). pandera: Statistical Data Validation of Pandas Dataframes. Accessed: 30 Aug 2021. URL: https://conference.scipy.org/proceedings/scipy2020/pdfs/niels_bantilan.pdf
- Di Zio, M., Fursova, N., Gelsema, T., Gießing, S., Guarnera, U., Petrauskienė, J., & Walsdorfer, K. (2018). Methodology for data validation 2.0. Essnet Validat Foundation, Brussels, Belgium, 1-85. Accessed: 13 Sept 2021. https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_methodology_for_data_validation_v2.0_-_rev2018_0.pdf
- Jeff Reback, jbrockmendel, Wes McKinney, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, gfyoung, Sinhrks, Matthew Roeschk, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Patrick Hoefler, Shahar Naveh, Marc Garcia, Jeremy Schendel, ... Kaiqi Dong. (2021). pandas-dev/pandas: Pandas 1.3.2 (v1.3.2). Zenodo. Accessed: 30 Aug 2021. https://doi.org/10.5281/zenodo.5203279
- Van der Loo, M. (2021) The Data Validation Cookbook version 1.0.4. Accessed: 30 Aug 2021. URL: https://data-cleaning.github.io/validate
- Van der Loo, M. (2021). The Data Validation Cookbook. Cran R Project. Accessed: 30 Aug 2021. https://cran.r-project.org/web/packages/validate/vignettes/cookbook.html
- Van der Loo, M. & de Jonge, E. (2020). Data Validation Infrastructure for R. Journal of Statistical Software, Accepted for publication. Accessed: 30 Aug 2021 https://arxiv.org/abs/1912.09759
- Van der Loo, M., & De Jonge, E. (2018). Statistical data cleaning with applications in R. John Wiley & Sons.
- Van der Loo, M. P. J., & de Jonge, E. (2021). Data Validation Infrastructure for R. Journal of Statistical Software, 97(10), 1–31. Accessed: 4 Oct 2021. https://doi.org/10.18637/jss.v097.i10

Islas Canarias
Del 15 al 19 de noviembre de 2021

¡Gracias!
Eskerrik asko!

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

Matias Ledesma Sereno: https://www.linkedin.com/in/matiasledesmasereno/
Ander Iparraguirre García: https://www.linkedin.com/in/ander-iparraguirre/